Scalable watermarking for identifying large language model outputs

Year: 2024 | *Citations: 150* | *Authors: Sumanth Dathathri, Abigail See, Sumedh Ghaisas*

Abstract

Large language models (LLMs) have enabled the generation of high-quality synthetic text, often indistinguishable from human-written content, at a scale that can markedly affect the nature of the information ecosystem1–3. Watermarking can help identify synthetic text and limit accidental or deliberate misuse4, but has not been adopted in production systems owing to stringent quality, detectability and computational efficiency requirements. Here we describe SynthID-Text, a production-ready text watermarking scheme that preserves text quality and enables high detection accuracy, with minimal latency overhead. SynthID-Text does not affect LLM training and modifies only the sampling procedure; watermark detection is computationally efficient, without using the underlying LLM. To enable watermarking at scale, we develop an algorithm integrating watermarking with speculative sampling, an efficiency technique frequently used in production systems5. Evaluations across multiple LLMs empirically show that SynthID-Text provides improved detectability over comparable methods, and standard benchmarks and human side-by-side ratings indicate no change in LLM capabilities. To demonstrate the feasibility of watermarking in large-scale-production systems, we conducted a live experiment that assessed feedback from nearly 20 million Gemini6 responses, again confirming the preservation of text quality. We hope that the availability of SynthID-Text7 will facilitate further development of watermarking and responsible use of LLM systems. A scheme for watermarking the text generated by large language models shows high text quality preservation and detection accuracy and low latency, and is feasible in large-scale-production settings.